Rethinking the Reference model in RLHF

letusgo126@126.com

March 2025

1 基于策略梯度方法的 PPO 奖励函数推导

1.1 RLHF 目标函数建模

基于人类反馈的强化学习（RLHF）的核心目标可建模为以下优化问题：

$$\arg\max_{\theta} J = \underbrace{\mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(y|x)}\left[r_\phi(x,y)\right]}_{\text{奖励最大化项}} - \beta \underbrace{\mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{\pi_\theta(\cdot|s_t,x)}\left[\mathbb{D}_{\mathrm{KL}}\left(\pi_\theta \,||\, \pi_{\mathrm{ref}}\right)\right]}_{\text{策略约束项}} \tag{1}$$

其中 β 为温度系数，控制策略偏离参考模型 π_{ref} 的程度。该目标函数包含两个关键部分：奖励期望的最大化和 KL 散度约束的平衡。

1.2 目标函数分解与分析

首先考虑奖励相关项的损失函数构造。根据策略梯度定理，可建立奖励项的损失函数为：

$$\mathcal{L}_R(\theta) = \mathbb{E}_{x\sim\mathcal{D},a_t\sim\pi_\theta}\left[r_\phi(x,y)\log\pi_\theta(a_t|s_t,x)\right] \tag{2}$$

结合 KL 散度约束项 $\mathcal{L}_{\mathrm{KL}_t}(\theta,\mathrm{ref})$，完整损失函数可表示为：

$$L_{total} = -\arg\min_{\theta} J = -\left(\mathcal{L}_R(\theta) - \beta\mathcal{L}_{\mathrm{KL}_t}(\theta,\mathrm{ref})\right) \tag{3}$$

1.3 梯度推导过程

对目标函数进行梯度分析时，需要分别处理两个组成部分：

1. **奖励项梯度**：

$$\nabla_\theta\mathcal{L}_R = \mathbb{E}_{x\sim\mathcal{D},a_t\sim\pi_\theta}\left[r_\phi(x,y)\nabla_\theta\log\pi_\theta(a_t|s_t,x)\right] \tag{4}$$

2. **KL 散度项梯度**（具体推导见引理 eq. (27)）：

$$-\nabla_\theta \mathcal{L}_{\mathrm{KL}_t} = -\mathbb{E}_{x\sim D}\mathbb{E}_{\pi_\theta}\left[\log\frac{\pi_\theta}{\pi_{ref}}\cdot\nabla_\theta\log\pi_\theta\right] \tag{5}$$

将二者结合可得总梯度：

$$\begin{aligned}\nabla_\theta J &= \nabla_\theta\mathcal{L}_R - \beta\nabla_\theta\mathcal{L}_{\mathrm{KL}_t}\\ &= \mathbb{E}_{x,a_t}\left[\left(r_\phi(x,y) - \beta\log\frac{\pi_\theta}{\pi_{ref}}\right)\nabla_\theta\log\pi_\theta\right]\end{aligned} \tag{6}$$

1.4 等效奖励函数构造

通过梯度分析可发现，原始优化问题可等价转换为：

$$\arg\max_\theta J' = \mathbb{E}_{x\sim\mathcal{D},a_t\sim\pi_\theta}\left[r_\phi(x,y) - \beta\log\frac{\pi_\theta(a_t|s_t,x)}{\pi_{ref}(a_t|s_t,x)}\right] \tag{7}$$

这揭示了 PPO 算法中奖励函数的设计本质：

$$r_t = r_\phi(x,y) - \beta\log\frac{\pi_\theta(a_t|s_t,x)}{\pi_{ref}(a_t|s_t,x)} \tag{8}$$

1.5 关键前提条件

需要特别强调的是，上述推导成立的关键在于 KL 散度项的梯度必须满足K1 及其类似的特定形式。当且仅当形式满足 k1 或类似形式时才成立。若该近似条件不成立，则公式 eq. (7) 和 eq. (8) 的推导过程将失效，因为 DPO 的推导基础 eq. (45) 就是该公式。

2 KL 惩罚项的数学性质分析

2.1 奖励函数中 KL 惩罚项的适用性条件

在 PPO 算法设计中，仅特定形式的 KL 散度估计量适合作为奖励函数的惩罚项。根据理论分析，k1 估计量具有数学适用性，其表达式如 eq. (8) 所示。然而，k2 和 k3 估计量由于内在的数学性质缺陷，不适合作为惩罚项，原因如下：

考虑 k2 和 k3 的估计值 kl_t 具有恒正特性，其梯度方向始终满足：

$$-\nabla_\theta \mathrm{KL}_t(\pi_\theta\|\pi_{ref}) = -kl_t\nabla_\theta\log\pi_\theta(a_t|s_t,x) \tag{9}$$

此时无论 $\pi_\theta(a_t|s_t,x)$ 与 $\pi_{ref}(a_t|s_t,x)$ 的概率分布关系如何，梯度更新方向都会强制降低当前策略 π_θ 生成任意动作 a_t 的概率。这种单向的惩罚机制本质上违背了策略优化算法的基本设计原则，即应建立方向可调的奖惩机制来引导策略改进。

2.2 KL 惩罚项的损失函数适配性

k2 估计量的有效性 如 eq. (28) 所示，k2 估计量能够构造有效的 KL 惩罚损失函数。其数学本质是通过 KL 散度的对称性设计，建立双向调节机制：当 π_θ 偏离 π_{ref} 时，梯度方向会根据偏离方向自动调整，既防止策略过度偏离，又保留必要的优化自由度。

k1 估计量的失效机理 k1 对应的损失函数形式为：

$$\mathcal{L}_{\mathrm{KL}_t}(\pi_\theta,\pi_{ref}) = \log \pi_\theta(a_t|s_t,x) - \log \pi_{ref}(a_t|s_t,x) \tag{10}$$

其梯度表达式揭示本质缺陷：

$$-\nabla_\theta \mathcal{L}_{\mathrm{KL}_t} = -\nabla_\theta \log \pi_\theta(a_t|s_t,x) \tag{11}$$

该梯度恒指向降低当前策略概率的方向，形成类似极大似然估计的反向约束。这种单向作用机制会导致策略网络的概率输出持续衰减，最终引发模型坍缩问题。

k3 估计量的近似特性 k3 构造的损失函数具有特殊形式：

$$\mathcal{L}_{\mathrm{KL}_t} = \frac{\pi_{ref}(a_t|s_t,x)}{\pi_\theta(a_t|s_t,x)} - \log \frac{\pi_{ref}(a_t|s_t,x)}{\pi_\theta(a_t|s_t,x)} - 1 \tag{12}$$

对应的梯度表达式揭示其近似本质：

$$-\nabla_\theta \mathcal{L}_{\mathrm{KL}_t} = \left(\frac{\pi_{\mathrm{ref}}}{\pi_\theta} - 1\right) \nabla_\theta \log \pi_\theta(a_t|s_t,x) \tag{13}$$

令 $x = \pi_{\mathrm{ref}}/\pi_\theta$，k2 与 k3 的梯度可对比表示为：

$$\text{k2 梯度}: \log x \cdot \nabla_\theta \log \pi_\theta \tag{14}$$

$$\text{k3 梯度}: (x-1) \cdot \nabla_\theta \log \pi_\theta \tag{15}$$

在策略邻域 $(x \approx 1)$ 进行泰勒展开时，$\log x \approx x - 1$，此时 k3 梯度构成 k2 梯度的线性近似。但这种近似具有两个关键缺陷：1. **有偏性**：当策略显

著偏离参考策略 (x 远离 1, 也就是训练后期 π_{ref} 离 π_θ 较远) 时，近似误差呈非线性增长 2. **非对称性**：$x-1$ 对 $\pi_\theta > \pi_{ref}$ 和 $\pi_\theta < \pi_{ref}$ 的响应特性不对称。

3 Diffusion RLHF

4 分类模型 KL 散度梯度推导

4.1 KL 散度的定义

KL 散度衡量两个离散概率分布 π_θ 和 π_{ref} 之间的差异，定义为：

$$\mathrm{KL}(\pi_\theta \| \pi_{ref}) = \mathbb{E}_{x\sim D, a_t \sim \pi_\theta(a_t|s_t,x)}\left[\log \frac{\pi_\theta(a_t|s_t,x)}{\pi_{ref}(a_t|s_t,x)}\right] = \mathbb{E}_{x\sim D}\sum_{a_t\in\mathcal{Y}} \pi_\theta(a_t|s_t,x)\log\frac{\pi_\theta(a_t|s_t,x)}{\pi_{ref}(a_t|s_t,x)}, \tag{16}$$

其中 $\mathcal{Y}$ 为离散类别空间。

4.2 KL 散度梯度推导步骤

步骤 1：展开 KL 表达式　直接写出离散求和形式：

$$\mathrm{KL}\big(\pi_\theta(\cdot|s_t,x)\|\pi_{ref}(\cdot|s_t,x)\big) = \sum_{a_t} \pi_\theta(a_t|s_t,x)\log\frac{\pi_\theta(a_t|s_t,x)}{\pi_{ref}(a_t|s_t,x)}. \tag{17}$$

步骤 2：应用梯度算子　对 θ 求梯度：

$$--\nabla_\theta \mathrm{KL}\big(\pi_\theta(\cdot|s_t,x)\|\pi_{ref}(\cdot|s_t,x)\big) = -\nabla_\theta \sum_{a_t} \pi_\theta(a_t|s_t,x)\log\frac{\pi_\theta(a_t|s_t,x)}{\pi_{ref}(a_t|s_t,x)}. \tag{18}$$

步骤 3：交换求和与梯度　由于求和项有限且 $\pi_\theta(y|x)$ 光滑，可交换求和与梯度：

$$-\nabla_\theta \mathrm{KL}\big(\pi_\theta(\cdot|s_t,x)\|\pi_{ref}(\cdot|s_t,x)\big) = -\sum_{a_t}\nabla_\theta\left[\pi_\theta(a_t|s_t,x)\log\frac{\pi_\theta(a_t|s_t,x)}{\pi_{ref}(a_t|s_t,x)}\right]. \tag{19}$$

步骤 4：乘积法则分解　对每一项应用乘积法则：

$$-\nabla_\theta\left[\pi_\theta(a_t|s_t,x)\log\frac{\pi_\theta(a_t|s_t,x)}{\pi_{ref}(a_t|s_t,x)}\right]=-[\underbrace{\nabla_\theta\pi_\theta(a_t|s_t,x)\cdot\log\frac{\pi_\theta(a_t|s_t,x)}{\pi_{ref}(a_t|s_t,x)}}_{\text{Term 1}}+\underbrace{\pi_\theta(a_t|s_t,x)\cdot\nabla_\theta\log\frac{\pi_\theta(a_t|s_t,x)}{\pi_{ref}(a_t|s_t,x)}}_{\text{Term 2}}]. \tag{20}$$

步骤 5：简化 Term 2　由于 $\pi_{ref}(y|x)$ 与 θ 无关，有：

$$\nabla_\theta\log\frac{\pi_\theta(a_t|s_t,x)}{\pi_{ref}(a_t|s_t,x)}=\nabla_\theta\log\pi_\theta(a_t|s_t,x)=\frac{\nabla_\theta\pi_\theta(a_t|s_t,x)}{\pi_\theta(a_t|s_t,x)}. \tag{21}$$

因此 Term 2 简化为：

$$\pi_\theta(a_t|s_t,x)\cdot\frac{\nabla_\theta\pi_\theta(a_t|s_t,x)}{\pi_\theta(a_t|s_t,x)}=\nabla_\theta\pi_\theta(a_t|s_t,x). \tag{22}$$

步骤 6：合并两项　将 Term 1 和 Term 2 相加：

$$\sum_{a_t}\left[\nabla_\theta\pi_\theta(a_t|s_t,x)\cdot\log\frac{\pi_\theta(a_t|s_t,x)}{\pi_{ref}(a_t|s_t,x)}+\nabla_\theta\pi_\theta(a_t|s_t,x)\right]. \tag{23}$$

步骤 7：处理归一化条件　由于 $\sum_{a_t}\pi_\theta(a_t|s_t,x)=1$，其梯度为 0：

$$\sum_{a_t}\nabla_\theta\pi_\theta(a_t|s_t,x)=\nabla_\theta\sum_{a_t}\pi_\theta(a_t|s_t)=\nabla_\theta 1=0. \tag{24}$$

因此第二项求和为 0，仅保留第一项：

$$-\nabla_\theta\mathrm{KL}\big(\pi_\theta(\cdot|s_t,x)\|\pi_{ref}(\cdot|s_t,x)\big)=-\sum_y\nabla_\theta\pi_\theta(a_t|s_t,x)\cdot\log\frac{\pi_\theta(a_t|s_t,x)}{\pi_{ref}(a_t|s_t,x)}. \tag{25}$$

步骤 8：对数导数技巧　利用 $\nabla_\theta\pi_\theta(a_t|s_t,x)=\pi_\theta(y|x)\nabla_\theta\log\pi_\theta(a_t|s_t,x)$，改写为：

$$-\nabla_\theta\mathrm{KL}\big(\pi_\theta(\cdot|s_t,x)\|\pi_{ref}(\cdot|s_t,x)\big)=-\sum_{a_t}\pi_\theta(a_t|s_t,x)\nabla_\theta\log\pi_\theta(a_t|s_t,x)\cdot\log\frac{\pi_\theta(a_t|s_t,x)}{\pi_{ref}(a_t|s_t,x)}. \tag{26}$$

步骤 9：期望形式 最终梯度可表示为期望：

$$
\begin{aligned}
-\nabla_\theta \mathrm{KL}_t(\pi_\theta \| \pi_{ref}) &= -\mathbb{E}_{x\sim D, a_t\sim\pi_\theta(a_t|s_t,x)}\left[\nabla_\theta \log \pi_\theta(a_t|s_t,x) \cdot \log \frac{\pi_\theta(a_t|s_t,x)}{\pi_{ref}(a_t|s_t,x)}\right] \\
&= \mathbb{E}_{x\sim D, a_t\sim\pi_\theta(a_t|s_t,x)}\left[\nabla_\theta \log \pi_\theta(a_t|s_t,x) \cdot \log \frac{\pi_{ref}(a_t|s_t,x)}{\pi_\theta(a_t|s_t,x})\right].
\end{aligned}
\tag{27}
$$

损失函数形式 根据梯度公式 eq. (27)，损失函数可推导为：

$$
\begin{aligned}
\mathcal{L}_{\mathrm{KL}_t}(\theta, \mathrm{ref}) &= \mathbb{E}_{x\sim D, a_t\sim\pi_\theta(a_t|s)}\left[\log \pi_\theta(a_t|s_t,x) \cdot \mathrm{SG}\left(\log \frac{\pi_\theta(a_t|s_t,x)}{\pi_{\mathrm{ref}}(a_t|s_t,x)}\right)\right] \\
&= \mathbb{E}_{x\sim D}\mathbb{E}_{a_t\sim\pi_\theta(a_t|s_t,x)}\frac{1}{2}\left[\log \frac{\pi_\theta(a_t|s_t,x)}{\pi_{\mathrm{ref}}(a_t|s_t,x)}\right]^2
\end{aligned}
\tag{28}
$$

其中 $\mathrm{SG}(\cdot)$ 表示阻断梯度（代码实现中对应 ‘.detach()‘ 函数）。

5 高维高斯分布的 KL 散度和梯度的推导

考虑两个 k 维各向同性高斯分布：$P \sim \mathcal{N}(\boldsymbol{\mu}_\theta, \sigma_1^2\mathbf{I})$, $Q \sim \mathcal{N}(\boldsymbol{\mu}_{\mathrm{ref}}, \sigma_2^2\mathbf{I})$ 其概率密度函数分别为：

$$
\begin{aligned}
P(\mathbf{x}) &= (2\pi\sigma_1^2)^{-k/2} \exp\left(-\frac{\|\mathbf{x}-\boldsymbol{\mu}_\theta\|^2}{2\sigma_1^2}\right) \\
Q(\mathbf{x}) &= (2\pi\sigma_2^2)^{-k/2} \exp\left(-\frac{\|\mathbf{x}-\boldsymbol{\mu}_{\mathrm{ref}}\|^2}{2\sigma_2^2}\right)
\end{aligned}
\tag{29}
$$

5.1 KL 散度定义

KL 散度定义为：

$$
D_{\mathrm{KL}}(P \parallel Q) = \mathbb{E}_{\mathbf{x}\sim P}\left[\ln \frac{P(\mathbf{x})}{Q(\mathbf{x})}\right] \tag{30}
$$

将密度函数代入，展开对数比：

$$
\ln \frac{P(\mathbf{x})}{Q(\mathbf{x})} = \frac{k}{2}\ln\frac{\sigma_2^2}{\sigma_1^2} + \left[-\frac{\|\mathbf{x}-\boldsymbol{\mu}_\theta\|^2}{2\sigma_1^2} + \frac{\|\mathbf{x}-\boldsymbol{\mu}_{\mathrm{ref}}\|^2}{2\sigma_2^2}\right] \tag{31}
$$

5.2 逐项计算期望

将 KL 散度拆分为常数项与二次项之和：

$$D_{\mathrm{KL}} = \underbrace{\frac{k}{2}\ln\frac{\sigma_2^2}{\sigma_1^2}}_{\text{常数项}} + \underbrace{\mathbb{E}_P\left[-\frac{\|\mathbf{x}-\boldsymbol{\mu}_\theta\|^2}{2\sigma_1^2} + \frac{\|\mathbf{x}-\boldsymbol{\mu}_{\mathrm{ref}}\|^2}{2\sigma_2^2}\right]}_{\text{二次项}} \tag{32}$$

计算 $\mathbb{E}_P[\|\mathbf{x}-\boldsymbol{\mu}_\theta\|^2]$ 由于 $\mathbf{x} \sim P$，协方差矩阵为 $\sigma_1^2\mathbf{I}$，故：

$$\mathbb{E}_P[\|\mathbf{x}-\boldsymbol{\mu}_\theta\|^2] = \mathrm{tr}(\sigma_1^2\mathbf{I}) = k\sigma_1^2 \tag{33}$$

计算 $\mathbb{E}_P[\|\mathbf{x}-\boldsymbol{\mu}_{\mathbf{ref}}\|^2]$ 展开二次型并取期望：

$$\begin{aligned}\|\mathbf{x}-\boldsymbol{\mu}_{\mathrm{ref}}\|^2 &= \|\mathbf{x}-\boldsymbol{\mu}_\theta+\boldsymbol{\mu}_\theta-\boldsymbol{\mu}_{\mathrm{ref}}\|^2 \\ &= \|\mathbf{x}-\boldsymbol{\mu}_\theta\|^2 + 2(\mathbf{x}-\boldsymbol{\mu}_\theta)^\top(\boldsymbol{\mu}_\theta-\boldsymbol{\mu}_{\mathrm{ref}}) + \|\boldsymbol{\mu}_\theta-\boldsymbol{\mu}_{\mathrm{ref}}\|^2\end{aligned} \tag{34}$$

取期望后，交叉项因 $\mathbb{E}_P[\mathbf{x}-\boldsymbol{\mu}_\theta]=0$ 而消失：

$$\mathbb{E}_P[\|\mathbf{x}-\boldsymbol{\mu}_{\mathrm{ref}}\|^2] = k\sigma_1^2 + \|\boldsymbol{\mu}_\theta-\boldsymbol{\mu}_{\mathrm{ref}}\|^2 \tag{35}$$

代入二次项

$$\begin{aligned}\mathbb{E}_P\left[-\frac{\|\mathbf{x}-\boldsymbol{\mu}_\theta\|^2}{2\sigma_1^2} + \frac{\|\mathbf{x}-\boldsymbol{\mu}_{\mathrm{ref}}\|^2}{2\sigma_2^2}\right] &= -\frac{k\sigma_1^2}{2\sigma_1^2} + \frac{k\sigma_1^2 + \|\boldsymbol{\mu}_\theta-\boldsymbol{\mu}_{\mathrm{ref}}\|^2}{2\sigma_2^2} \\ &= -\frac{k}{2} + \frac{k\sigma_1^2}{2\sigma_2^2} + \frac{\|\boldsymbol{\mu}_\theta-\boldsymbol{\mu}_{\mathrm{ref}}\|^2}{2\sigma_2^2}\end{aligned} \tag{36}$$

合并所有项 将常数项与二次项合并：

$$\begin{aligned}D_{\mathrm{KL}} &= \frac{k}{2}\ln\frac{\sigma_2^2}{\sigma_1^2} - \frac{k}{2} + \frac{k\sigma_1^2}{2\sigma_2^2} + \frac{\|\boldsymbol{\mu}_\theta-\boldsymbol{\mu}_{\mathrm{ref}}\|^2}{2\sigma_2^2} \\ &= \frac{k}{2}\left[2\ln\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2}{\sigma_2^2} - 1\right] + \frac{\|\boldsymbol{\mu}_\theta-\boldsymbol{\mu}_{\mathrm{ref}}\|^2}{2\sigma_2^2}\end{aligned} \tag{37}$$

最终公式 整理后得到：

$$D_{\mathrm{KL}}(P \,\|\, Q) = \underbrace{k\left(\ln\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2}{2\sigma_2^2} - \frac{1}{2}\right)}_{\text{方差项}} + \underbrace{\frac{\|\boldsymbol{\mu}_\theta-\boldsymbol{\mu}_{\mathrm{ref}}\|^2}{2\sigma_2^2}}_{\text{均值项}} \tag{38}$$

当 $\sigma = \sigma_1 = \sigma_2$ 时，

$$D_{\mathrm{KL}}(P \,\|\, Q) = \frac{\|\boldsymbol{\mu}_\theta-\boldsymbol{\mu}_{\mathrm{ref}}\|^2}{2\sigma^2} \tag{39}$$

5.3 对 $\boldsymbol{\mu}_\theta$ 的梯度推导

从 KL 散度公式中提取与 $\boldsymbol{\mu}_\theta$ 相关的项：

$$D_{\mathrm{KL}}(P \parallel Q) = \underbrace{\frac{\|\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\mathrm{ref}}\|^2}{2\sigma_2^2}}_{\text{唯一与 } \boldsymbol{\mu}_\theta \text{ 相关的项}} + \text{其他与 } \boldsymbol{\mu}_\theta \text{ 无关的项}. \tag{40}$$

步骤 1：展开二次型 对 $\|\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\mathrm{ref}}\|^2$ 展开：

$$\|\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\mathrm{ref}}\|^2 = (\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\mathrm{ref}})^\top (\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\mathrm{ref}}). \tag{41}$$

步骤 2：计算关于 $\boldsymbol{\mu}_\theta$ 的梯度 利用二次型的梯度公式：

$$-\nabla_{\boldsymbol{\mu}_\theta} \left[(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\mathrm{ref}})^\top (\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\mathrm{ref}}) \right] = -2(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\mathrm{ref}}). \tag{42}$$

步骤 3：组合梯度分量 将梯度结果代入 KL 散度表达式：

$$-\nabla_{\boldsymbol{\mu}_\theta} D_{\mathrm{KL}} = -\frac{1}{2\sigma_2^2} \cdot 2(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\mathrm{ref}}) = -\frac{\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\mathrm{ref}}}{\sigma_2^2}. \tag{43}$$

最终梯度表达式

$$-\nabla_{\boldsymbol{\mu}_\theta} D_{\mathrm{KL}}(P \parallel Q) = -\frac{\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\mathrm{ref}}}{\sigma_2^2} \tag{44}$$

6 DPO 推导

本节基于 KL 约束的奖励最大化目标，推导出可操作的直接偏好优化目标函数。首先建立基础优化问题：

$$\begin{aligned}
&\max_\pi \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi}[r(x,y)] - \beta D_{\mathrm{KL}}[\pi(y|x) \| \pi_{\mathrm{ref}}(y|x)] \\
&= \max_\pi \mathbb{E}_{x\sim\mathcal{D}} \mathbb{E}_{y\sim\pi(y|x)} \left[r(x,y) - \beta \log \frac{\pi(y|x)}{\pi_{\mathrm{ref}}(y|x)} \right] \\
&= \min_\pi \mathbb{E}_{x\sim\mathcal{D}} \mathbb{E}_{y\sim\pi(y|x)} \left[\log \frac{\pi(y|x)}{\pi_{\mathrm{ref}}(y|x)} - \frac{1}{\beta} r(x,y) \right]
\end{aligned} \tag{45}$$

为求解该优化问题，引入配分函数 $Z(x)$ 构造概率分布：

$$Z(x) = \sum_y \pi_{\mathrm{ref}}(y|x) \exp\left(\frac{1}{\beta} r(x,y) \right) \tag{46}$$

定义新的参考分布 π^* 为：

$$\pi^*(y|x) = \frac{1}{Z(x)}\pi_{\text{ref}}(y|x)\exp\left(\frac{1}{\beta}r(x,y)\right) \tag{47}$$

将目标函数重构为：

$$\min_{\pi} \mathbb{E}_{x\sim\mathcal{D}}\left[D_{\text{KL}}(\pi(y|x)\|\pi^*(y|x)) - \log Z(x)\right] \tag{48}$$

由于 $Z(x)$ 与策略 π 无关，优化目标简化为最小化 KL 散度项。根据 KL 散度的非负性，当 $\pi = \pi^*$ 时取得全局最优解。

6.1 从奖励建模到偏好学习

实际应用中直接求解 π^* 存在两大障碍：1) 真实奖励函数 r^* 未知；2) 配分函数 $Z(x)$ 的计算涉及全响应空间积分。为此，我们引入偏好学习框架。

采用 Bradley-Terry 模型，对于输入 x 和响应对 (y_w, y_l)，偏好概率建模为：

$$p^*(y_w \succ y_l|x) = \frac{\exp(r^*(x,y_w))}{\exp(r^*(x,y_w)) + \exp(r^*(x,y_l))} \tag{49}$$

关键突破在于建立奖励函数与最优策略的显式关联。由式 (2) 可得：

$$r(x,y) = \beta\log\frac{\pi^*(y|x)}{\pi_{\text{ref}}(y|x)} + \beta\log Z(x) \tag{50}$$

将奖励差表达式代入偏好概率模型：

$$p^*(y_w \succ y_l|x) = \frac{1}{1+\exp\left(\beta\log\frac{\pi^*(y_l|x)}{\pi_{\text{ref}}(y_l|x)} - \beta\log\frac{\pi^*(y_w|x)}{\pi_{\text{ref}}(y_w|x)}\right)} \tag{51}$$

6.2 最终目标函数

通过极大似然估计，得到直接优化策略的参数化目标函数：

$$\mathcal{L}_{\text{DPO}}(\pi_\theta;\pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right] \tag{52}$$

该目标函数具有三个显著优势：1) 避免显式奖励建模；2) 规避配分函数计算；3) 可直接通过策略网络进行梯度优化。实验表明，这种参数化形式能有效对齐模型输出与人类偏好。

7 重要性采样

loss 中由 $\log \pi_\theta(a_t|s_t)$ 转为 $\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$

$$\begin{aligned}
\nabla_\theta J(\theta) &= \mathbb{E}_{(s_t,a_t)\sim\pi_\theta}\left[A_\theta(s_t,a_t)\nabla_\theta \log \pi_\theta(a_t|s_t)\right] \\
&= \mathbb{E}_{(s_t,a_t)\sim\pi_{\theta'}}\left[\frac{\pi_\theta(s_t,a_t)}{\pi_{\theta'}(s_t,a_t)}A_\theta(s_t,a_t)\nabla_\theta \log \pi_\theta(a_t|s_t)\right] \\
&= \mathbb{E}_{(s_t,a_t)\sim\pi_{\theta'}}\left[\frac{\pi_\theta(a_t|s_t)\pi_\theta(s_t)}{\pi_{\theta'}(a_t|s_t)\pi_{\theta'}(s_t)}A_\theta(s_t,a_t)\nabla_\theta \log \pi_\theta(a_t|s_t)\right] \\
&\approx \mathbb{E}_{(s_t,a_t)\sim\pi_{\theta'}}\left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta'}(a_t|s_t)}A_{\theta'}(s_t,a_t)\nabla_\theta \log \pi_\theta(a_t|s_t)\right] \\
&= \mathbb{E}_{(s_t,a_t)\sim\pi_{\theta'}}\left[\frac{\nabla\pi_\theta(a_t|s_t)}{\pi_{\theta'}(a_t|s_t)}A_{\theta'}(s_t,a_t)\right]
\end{aligned} \tag{53}$$

Let $\theta' = \theta_{\text{old}}$, the off policy objective can be expressed as:

$$J(\theta) = \mathbb{E}_{(s_t,a_t)\sim\pi_{\theta_{\text{old}}}}\left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}A_{\theta_{\text{old}}}(s_t,c_{\text{可}}\right] \tag{54}$$